VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2017, ISSN(e): 2309-6519; ISSN(p): 2411-6327
Volume 12, Number 3, September-December, 2017 *pp.43-48*

EXTRACTING EVENTS FROM SOCIAL MEDIA USING NLP

SAWABIA NASEEM
` Department of Computer Sciences, University of Management and Technology, Lahore, Pakistan
Email: F2017279007@umt.edu.pk

ABSTRACT. *Now a day's Social media is major channel of communication between individuals and organizations. Huge data is available over the social networks, so it is important and essential to analyze this data to extract information. The data on social media is very much scattered, to extract information it needs to be organized. Natural Language Processing (NLP) techniques are used to analyze the scattered data to fetch information for targeted entities (Event, Category, Date, Place, and Time period).Extracted information it is listed on database and can be used in several ways. In this paper, a model is proposed which categorize event by their types, Date Place and Time. The results show this model can categorize the 90% events.*
Keywords*:* NLP, Scrapping, POS, Chunking, Chinking, Tokenization.

1. **Introduction.** There is no doubt that social media change our world's traditional communication style dramatically [1]. It has a great impact at all fields of life. At social media every day events are sponsored related to different categories. People have no idea about all events thoroughly so they need to search all events one by one. This task is so hectic to perform it manually, and store all required data at note book or calendar. A lot of posts are post on Facebook on daily basis, having information about events, including Date, Time, Place and time period for which this event is valid. In this world of technology there is a need to analyze this huge data to summarize the required information. For this purpose data mining techniques are used to make this information user-friendly[2][3]. NLP is a technique which deals with natural languages. The work on NLP started since 1980, Alan turning writes an article on Turing sets. It is most frequently using technique for data mining now a day's. For this purpose first summarizes the data, perform chunking and tagging on data and then reorganization process is performed. It performs extraction according to the user requirement and makes a bridge for computer and application to understand human language.

This proposed model use NLP technique to extract events from social media and categorize them on the basis of time, place, city and category (entertainment, education, political). For this purpose, select keyword related to each event and collects all data against every keyword from Facebook; perform scraping at HTML for type, date, place and time period. In fig 1.1 summarize categorization process is given bellow in a simple flow chart.

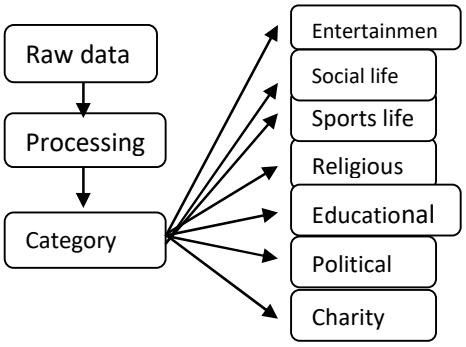

Fig. 1.1 flow diagram of categorization process

2. Related Work. Natural language processing[3] is a language which is used as a bride between computer and human languages. By using this language computers precede a large number of human languages. The work on NLP started since 1980, Alan turning writes an article on Turing sets.

Miler et al, was the first person who proposed the parsing of a text to extract information [1997].They explain the syntactic explanations of the Penn Treebank corpus (Marcus et al)[1993] with data and relation mentions particular to the MUC-7 assessment (Chinchor et al)[1997] example, worker of associations that hold in the middle of individual and organization called entities and after that prepared to generate a parsing over this joined semantic representation. In a similar way, Finke Keeping an eye on (2009) blended the syntactic and proposed a notation, named entity of the OntoNotes corpus (Hovy et al)[2006] and prepared a discriminative parsing model for the joint issue of syntactic parsing and named element acknowledgment. As both methods are require a same notation of syntactic.

Also, semantic components, which isn't generally feasible, and concentrated just on named parallel entities. In proposed model, there is no need to focus on entity; information is extracted from the HTML source directly. Finkel and Manning (2009b) likewise proposed a parsing model for the extraction of advanced named entities, which, similar to proposed work, parses only the required word or semantic.

3. Methodology. This paper proposed a model[19] which categorizes events sponsored on social media on the basis of type, time, place and valid time period. The process is explained in detail below (figure 1.5). The input is given in form of HTML which is copied against the sponsored event, and output will be generated in the form of list holding corresponding data about all events.

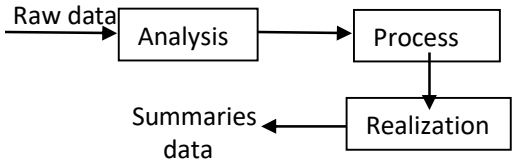

Fig 1.2 Detail of summarization process

Summarization of data. Summarization of data is a process[3][11] in which data is summarized to collect all the required data set from complete document (fig 1.2).

This process normally used two approaches.

Extraction and Abstraction. Extraction is a process in which words form a subset is selected from the original text to form a summery. Relatively in abstraction, semantic keywords are used to summarize the text to extract information.

Example of summarization using [X]

Input (x1,…,Xn).

Pakistan defense minister called force at Sunday for the creation of a joint front for combating global terrorism

Output (y1,…,ys).

Pakistan calls for joint front against terrorism⇐ g(terrorism, x, for, joint, front, against)
Fig 1.3 algorithm for sumrization process
Tokenization. The original text is splits[19] in to Tokens by using split() function. Split function(fig. 1.4) take text as input and then convert it into tokens by checking bank spaces and periods[5][2].

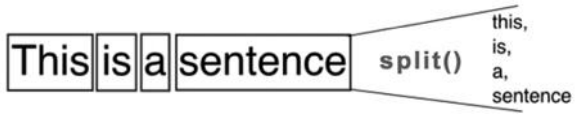

Fig. 1.4 n-gram algorithm example

In fig 1.4 tokenization processes is perform by using an example. After tokenization this data passed to Tagging.[2]

Tagging. Every language has its own part of speech. Different parts of speech could be added in a dictionary according to date, event, place etc. For this purpose it is necessary to define a list. Stanford Part of speech Tagger is used to for tagging [13].

Table:1 Parts of speech(POS) Tagging

Words	Tags
Noun	NN
Verb	VB
Conjunction	CON
Month	month
Year	year
Punctuation	PUN
Expression	EXP
Modal	MDL
Preposition	IN
Date	DAT
Education	EDU
Entertainment	ENT
Place	PLC
Time Period	TP
Category	CTG
Lahore	LHR
Seminar	SR

An example is given below,
"Today is a sports event"
POS tagging
{('today','DATSPC'),('is','VB'),('a','DT'),('sports','NN'),('event','NN')}

In this example, a sentence is converting into tags by using POS tagging process. Each word tag according its tag words explain in the tag-set in the model.

Chunking & Chinking. After POS tagging, it is not necessary we get accurate information according to our task. To solve this problem, two techniques are used, chunking and chinking. The selection of tags which are regarding our needs and will be proceed is called chunking. The tags which are not required for our task are called chinking and these words eliminated from the list.

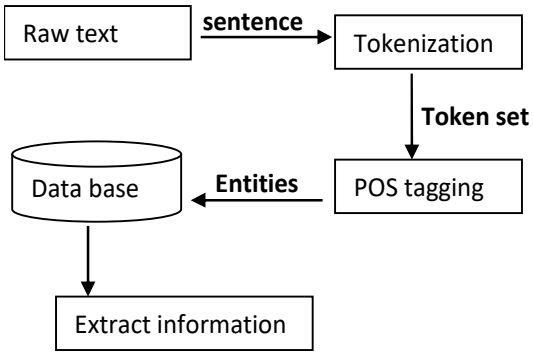

Fig:1.5 process of event extraction

Event Extraction

The event sponsored on Facebook is of different types containing information about Date, Time, Place, validity time period. The event extraction process is divided into five parts, category, date, place, time, validity of time [4][5].

Programming code for scraping is given bellow.

```
serp = browser.go_to facebook_event_url
serp_html = serp.html
only_events = serp.css('#id_event_div').map{|ev| {:name => ev.css('#rds_name'), :date => ev.css('#rsa_ddt'), :description => ev.css('#_dds')}}
all_results = only_events
```

Fig 1.6 scraping HTML against each keyword

Above fig 1.6 explain the algorithm of scraping which is performed on data which is collected in HTML form against each keyword from social media. This algorithm gives only “events” in result section and all results are in form of array of hashes.

```
Sample data form = {:name, :date, :description}
Perform nlp related operations using nlp gem
Categories = []
all_results.each do |ar|
categoriese << {:event => ar, :category => NLP.categorise(ar)}
end
```

Fig 1.7 Process scrap-data using NLP

Above figure describe the algorithm of extraction process of date, time, place and category of event.

(a) Subject Extraction. The event is held for a specific subject. The sponsored post can b sponsor with multiple words or adjacent words, so every related word will be processed to get information. If words are noun, conjunction or number it will be included in subject.

(b) Place Extraction. Most of the time, Place of an event is found after the subject and after the words “in” and “at” .All these words are put into the array and extraction is performed for each word to extract the place.

(c) Date Extraction. Date is normally written in a standard form DD/MM/YYYY. Date is written in any format is convert into this standard format.

Most of the time date is written into un-usual ways like, 3rd April, Today, next Monday. In this case date is extract by using different methods. For example, to extract date of “next Monday”, date of today is get and count till next Monday. To performing date extraction, date module is used timedelta () by using python.

```
Print: timedelta(days=365,hours=8,mintues=18)
Time:    365 days 8:18:00
```

(d) Time extraction. Time can likewise be determined from multiple points of view like 930 am/pm,9.30 AM/PM,9, and so on time which have am or pm can be straightforwardly distinguished from the message and can be added to the time. Yet, for alternate cases the digits showing up after words like at, from, after, and so on are added to time. As default “AM” is given to the time which doesn’t have the “am” or “pm” part.

Print: timedelta(days=365,hours=8,mintues=18)
Time: 365 days 8:18:00
(e) Validation. It is the procedure were the time period of validation is checked for the event. On the off chance that an occasion contain the subject and some other part then they are taken as a substantial occasion. In any case, there can be cases were the date and time separated can be invalid. To check for date legitimacy, date acquired are passed to check_date() work which return 1 in the event that it is valid and 0 on the off chance that it isn't. On the off chance that date is invalid the date part is set are invalid.

Results. Some samples are given bellow to describe the results of this proposed model .

Sample 1:
Children learning expo in expo center Lahore.
Results Sample 1
: [{:name => 'Education Expo Lahore', :category => 'Education'}, {:name => 'Children learning', :category => 'Education'}, {:place => ‘expo center’]
Category: Education
Name: Children learning
Place: Expo center

Sample 2:
Art exhibition, 7th feb, expo center.
Result sample 2:
[{:name => ‘Art exhibition, 7th feb, expo center’, :category => ‘Art’}]
Category: Exhibition
Date: 7th feb
Place: expo center
Name : Art exhibition

Sample 3: Lahore international book fair ,feb 5, Gulberg111.
Result sample 3:
[{:name => ‘Lahore international book fair ,feb 5, Gulberg111’, :category => ‘Education’}]
Category: Exibition
Place: Gulberg111
Date: 5 feb
Name: Book fair

Sample 4: Productive muslima ,mar 15, Alnoor international, jahar town Lahore.
Result sample 4:
[{:name => ‘Productive muslima ,mar 15, Alnoor international, jahar town Lahore’, :category => ‘Islam’}]
Category: Islamic
Place: johar town
Date: 15 mar
Name: Productive muslima

Conclusion. This proposed model can categorize[20] different events sponsored on Facebook at daily basis, and make a database at the backend by scraping the html of each event. Each event will be extract by giving category name and all corresponding information will be extract automatically. This model can be upgrade further for different tasks by using other techniques. Simply it makes easy to search an event with all its detail and saved in a backend Data set for further use than search each event manually.

REFERENCES

[1] Dalli, A. (2004). Automated email integration with personal information management application. *The UK special-interest group for computational linguistics*.

[2] Virmani, C., Pillai, A., & Juneja, D. (2017). Extracting Information from Social Network using NLP. *International Journal of Computational Intelligence Research*, *13*(4), 621-630.

[3] Dalli, A. (2004). Automated email integration with personal information management application. *The UK special-interest group for computational linguistics*.

[4] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

[5] Tokenization, https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html, 24-Feb-2018

[6] Bharti, S. K., & Babu, K. S. (2017). Automatic keyword extraction for text summarization: A survey. *arXiv preprint arXiv:1704.03242*.

[7] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107). Association for Computational Linguistics.

[8] Seon, C. N., Kim, H., & Seo, J. (2011). Efficient appointment information extraction from short messages in mobile devices with limited hardware resources. *Pattern Recognition Letters*, *32*(2), 127-133.

[9] Cooper, R., Ali, S., & Bi, C. (2005, June). Extracting information from short messages. In *International Conference on Application of Natural Language to Information Systems* (pp. 388-391). Springer, Berlin, Heidelberg.

[10] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[11] The Stanford Natural Language Processing, Stanford Named Entity recognize"Available: http://nlp.stanford.edu/software/CRF-NER.shtml#About.2012

[12] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[13] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

[14] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493-2537.

[15] Hogenboom, F., Frasincar, F., Kaymak, U., & De Jong, F. (2011, October). An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)* (Vol. 779, pp. 48-57). Koblenz, Germany: CEUR-WS. org.

[16] Horecki, K., & Mazurkiewicz, J. (2015, June). Natural Language Processing Methods Used for Automatic Prediction Mechanism of Related Phenomenon. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 13-24). Springer, Cham.

[17] Surabhi, M. C. (2013, July). Natural language processing future. In *Optical Imaging Sensor and Security (ICOSS), 2013 International Conference on* (pp. 1-3). IEEE.

[18] Abid, A., Hussain, N., Abid, K., Ahmad, F., Farooq, M. S., Farooq, U., ... & Sabir, N. (2016). A survey on search results diversification techniques. *Neural Computing and Applications*, *27*(5), 1207-1229.

[19] Adnan Abid, Muhammad Shoaib Farooq, and Ishaq Raza, Variants of Teaching First Course in Database Systems, Bulletin of Education and Research, December 2015, Vol. 37, No. 2 pp. 1--17.

[20] Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., ... & Varela, I. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine*, *366*(10), 883-892.